NGHIÊN CỨU GIẢI PHÁP XÂY DỰNG HỆ THỐNG TỔNG HỢP VÀ HỖ TRỢ TƯ VẤN VIỆC LÀM

SOLUTIONS FOR BUILDING A SYSTEM SUPPORTING THE INFORMATION SYNTHESISAND CONSULTANCY OF CAREERS

Trần Thị Kiều, Nguyễn Văn Bình, Huỳnh Công Pháp

Trường Cao đẳng Công nghệ Thông tin, Đại học Đà Nẵng
Email: thkieu105@gmail.com, binhsht@gmail.com, hcphap@gmail.com

Tóm tắt - Hiện nay nhu cầu về tìm kiếm cũng như giới thiệu việc làm ngày càng tăng rõ rệt. Để đáp ứng nhu cầu đó, có rất nhiều trang web khác nhau giới thiệu và hỗ trợ tìm kiếm việc làm. Với sự tồn tại quá nhiều trang web về việc làm như vậy đã dẫn đến một thực trạng là thông tin nằm rải rác, rời rạc và nhiều khi trùng lặp ở các trang web khác nhau, làm cho người dùng tốn rất nhiều thời gian và công sức để tìm ra những thông tin mà mình quan tâm cũng như nhìn thấy bức tranh toàn diện về việc làm. Trước thực trạng đó, chúng tôi đề xuất xây dựng một hệ thống tổng hợp thông tin việc làm bằng cách thu thập tự động thông tin từ các website việc làm khác nhau. Từ đó biểu diễn lại thông tin việc làm một cách hệ thống, nhằm giúp người dùng dễ dàng hơn trong việc xem và tìm kiếm thông tin việc làm. Đồng thời từ thông tin tổng hợp được, hệ thống sẽ thống kê việc làm theo nhiều tiêu chí khác nhau như vị trí công việc, địa bàn, thời gian, …, nhằm giúp cho người dùng nhìn thấy tổng quan hơn về tình hình của thị trường lao động cũng như phục vụ cho việc tư vấn việc làm.

Từ khóa - trích xuất thông tin; phân loại văn bản; DOM Tree; SVM; tư vấn việc làm.

Abstract - The present demand for job finding as well as job introduction is on a dramatic increase. To meet this demand, there are many different career websites. With so many career sites, information become scattered, disjointed and often duplicated in different sites, taking a lot of time and effort to find the information that interests you as well as see the comprehensive picture of employment. In this situation, we look forward to building a system supporting the information synthesis and consultancy of careers by retrieving the information from different career websites. In addition, our system shows the job information in a systematic way, in order to facilitate users to view and search the job postings. Also, the system classifies information for easier search based on the criteria such as profession, area and time... to help users see a more general situation of labor market as well as consulting service jobs.

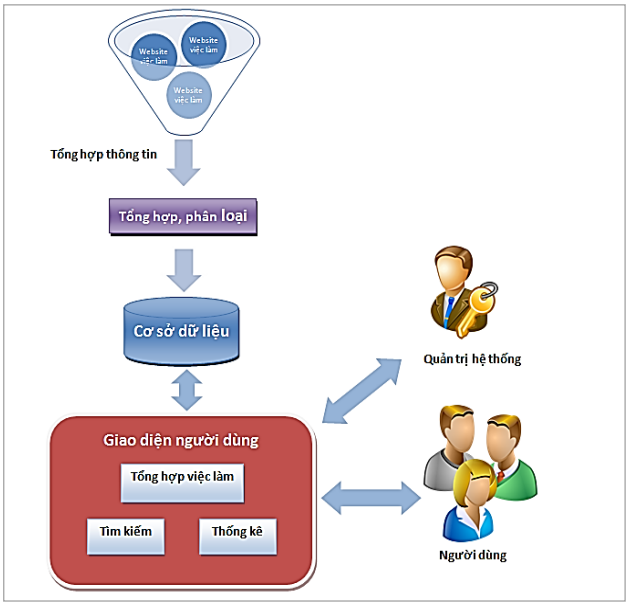
Key words - information extraction; text classification, DOM tree; SVM; consultancy of careers.

1. Đặt vấn đề

Hiện nay, nhu cầu về tìm kiếm việc làm rất lớn. Theo thống kê năm 2013 về tỉ lệ thất nghiệp ở Việt Nam: tỷ lệ thất nghiệp tăng từ 1,81% lên 1,9% (so với cùng kỳ năm 2012), trong đó tỷ lệ thất nghiệp thanh niên (từ 15-24 tuổi) tăng từ 5,29% tới 5,95% (so với cùng kỳ năm 2012) [21]. Nhu cầu về tìm kiếm, giới thiệu việc làm cũng như đăng tin tuyển dụng ngày càng tăng rõ rệt. Người đi tìm việc cũng như việc đi tìm người thường thông qua các kênh thông tin, các phương tiện như Internet.

Hiện tại có rất nhiều trang thông tin việc làm, cụ thể như: vietnamworks.com, timviecnhanh.com, careerlink.com,… Các nhà tuyển dụng có thể đăng tin tuyển dụng lên nhiều trang web dẫn tới tình trạng trùng lặp thông tin giữa các trang. Người đi tìm việc trước hết phải biết được địa chỉ của các trang hay phải lần mò từng trang để tìm ra việc làm thích hợp với điều kiện và năng lực của mình. Hơn nữa, thông tin việc làm rời rạc, không có hệ thống.

Với cách lưu trữ thông tin như vậy, không giúp cho người đi tìm việc nhìn thấy được bức tranh toàn diện về việc làm. Chẳng hạn như: không thấy được nhu cầu việc làm hiện nay như thế nào, nhu cầu về ngành nghề hiện nay ra sao hay không thấy con số về việc làm cũng như không có khả năng tổng hợp thông tin việc làm. Điều đó làm cho người đi tìm việc bối rối giữa đống bùi nhùi các công việc phải chọn lựa, làm tiêu tốn không ít thời gian và công sức. Hay các trường cao đẳng, đại học hiện nay không định hướng được việc làm để đào tạo, hay những người học không có cơ sở thực sự để chọn cho mình một ngành nghề trong tương lai.



Hình 1. *Mô hình bài toán*

Với những thực trạng đó, chúng tôi mong muốn xây dựng một hệ thống tổng hợp các thông tin việc làm bằng cách thu thập các nguồn thông tin từ các website việc làm để xây dựng một website có hệ thống thông tin việc làm một cách tổng hợp và phân loại thông tin phục vụ cho việc tìm kiếm dễ dàng hơn theo các tiêu chí: ngành nghề, khu vực và thời gian. Từ thông tin tổng hợp này, thống kê những số liệu phục vụ cho việc dự đoán xu hướng, định hướng và tư vấn việc làm. Chẳng hạn như: thống kê tỷ lệ

việc làm hiện nay tăng hay giảm cũng như dự đoán được giai đoạn sắp tới ngành nghề nào được tuyển nhiều nhất với tỷ lệ tăng nhanh nhất,… Và đây cũng có thể là một diễn đàn cho phép người tuyển dụng đưa thông tin, cũng là nơi gặp gỡ giữa những người lao động và sử dụng lao động.

2. Mô hình tổng quát bài toán

Hình 1 là mô hình tổng quát các bước xây dựng hệ thống tổng hợp thông tin và hỗ trợ tư vấn việc làm. Thông tin về việc làm được đăng tải rải rác trên các hệ thống website việc làm sẽ được thu thập một cách tự động để xử lý. Sau khi được xử lý, dữ liệu được phân loại và lưu trữ vào cơ sở dữ liệu nhằm phục vụ việc hiển thị cũng như thống kê.

3. Tổng quan và lựa chọn giải pháp kỹ thuật

3.1. Trích xuất thông tin

Trích xuất thông tin (Information Extraction – IE) chính là bóc tách các thông tin "có ích" đối với người dùng từ website.

Dữ liệu thông thường được chia thành 3 dạng cơ bản [17]:

- Dữ liệu không cấu trúc.
- Dữ liệu có cấu trúc.
- Dữ liệu bán cấu trúc.

Các trang web thông thường là một dạng tiêu biểu của dữ liệu bán cấu trúc.

3.1.1. Các hướng tiếp cận trong bài toán trích xuất thông tin đối với dữ liệu bán cấu trúc

Hiện nay, crawler hay wrapper [20] thường được sử dụng để bóc tách nội dung trên web. Một wrapper có thể được xem như là một thủ tục được thiết kế để có thể trích xuất được những nội dung cần quan tâm từ một nguồn thông tin nào đó. Hiện nay, trên thế giới đã có nhiều công trình nghiên cứu khác nhau, sử dụng nhiều phương pháp tạo wrapper khác nhau để thực hiện trích xuất thông tin trên web. Có thể liệt kê một số phương pháp sau đây:

+ Phân tích mã HTML

VietSpider [16] của tác giả Nhữ Đình Thuần: hệ thống xử lý nội dung định dạng theo HTML để chuyển đổi chúng về mô hình dữ liệu dạng tree mà chính xác là Tree DOM. Dựa và Tree DOM, có thể dễ dàng truy xuất những thành phần của nó.

Có nhiều phương pháp phân tích mã HTML: phương pháp thủ công, phương pháp wrapper quy nạp, phương pháp tự động [6].

+ Phương pháp so trùng

Ứng dụng thuật toán phân lớp trích xuất thông tin văn bản FSVM trên Internet [1] của tác giả Vũ Thanh Nguyên, Trang Nhật Quang: Trích xuất thông tin bằng cách so trùng hai trang web. Phương pháp này được thực hiện bằng cách so trùng trang web cần trích xuất với một trang web mẫu để xác định khung trình bày chung của hai trang web, từ khung trình bày chung ta có thể trích xuất ra được nội dung chính của trang web cần rút trích.

+ Xử lý ngôn ngữ tự nhiên

Đây là phương pháp sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên được áp dụng cho những tài liệu mà thông tin trên đó thường không có một cấu trúc nhất định (như truyện). Các kỹ thuật này xem xét sự ràng buộc về mặt cú pháp và ngữ nghĩa để nhận dạng ra các thông tin liên quan và rút trích ra thông tin cần thiết cho các bước xử lý nào đó. Các công cụ sử dụng phương pháp này thích hợp cho việc rút trích thông tin trên những trang web có chứa những đoạn văn tuân theo quy luật văn phạm. Một số công cụ sử dụng phương pháp xử lý ngôn ngữ tự nhiên trong việc bóc tách nội dung như: WHISK hay RAPIER.

3.1.2. Phương pháp trích xuất thông tin cho dữ liệu bán cấu trúc

a. Phân tích mã HTML

Có nhiều phương pháp phân tích mã HTML: phương pháp thủ công, phương pháp wrapper quy nạp, phương pháp tự động [6].

- Hiện nay ý tưởng của phương pháp trích xuất thủ công không còn được sử dụng.

- Phương pháp wrapper quy nạp lại phụ thuộc vào việc gán nhãn bằng tay nên nó không phù hợp cho việc trích xuất một số lượng lớn các trang. Nếu một trang web tổng hợp việc làm muốn trích xuất tất cả các việc làm từ các website, việc gán nhãn bằng tay hầu như là không thể. Việc duy trì wrapper là việc làm rất tốn kém, vì web là một môi trường động. Các site thì luôn luôn thay đổi.

- Phương pháp tự động: Được đề xuất trong năm 1998, phương pháp này tự động tìm các mẫu hoặc các cấu trúc để trích xuất thông tin từ những trang cho trước. Vì phương pháp này không cần đến sự gán nhãn bằng tay nên nó có thể trích xuất được dữ liệu từ một lượng khổng lồ các trang; một số giải thuật tiêu biểu như RoadRunner, bootstrapping. Phương pháp này đã khắc phục được những nhược điểm của wrapper quy nạp. Việc trích xuất tự động là hoàn toàn có thể bởi vì dữ liệu trên một website thường được mã hóa với một số lượng mẫu cố định. Có thể tìm những khuôn mẫu đó bằng việc khai phá những mẫu lặp lại trong nhiều trang của một website.

Và một thuật toán trích xuất tự động khá tiêu biểu đó là RoadRunner [15].

- Ưu, nhược điểm của giải thuật:

Ưu điểm: Không cần sự gán nhãn của người dùng với tập mẫu huấn luyện, có thể tự động xây dựng được mẫu trích xuất.

Nhược điểm: Nó không thể tự động nhận dạng được đâu là thực thể thông tin mong muốn của người dùng. Vì vậy, người sử dụng sẽ vẫn phải tự gán nhãn những kết quả đầu ra.

b. Trích xuất thông tin dựa vào cây DOM

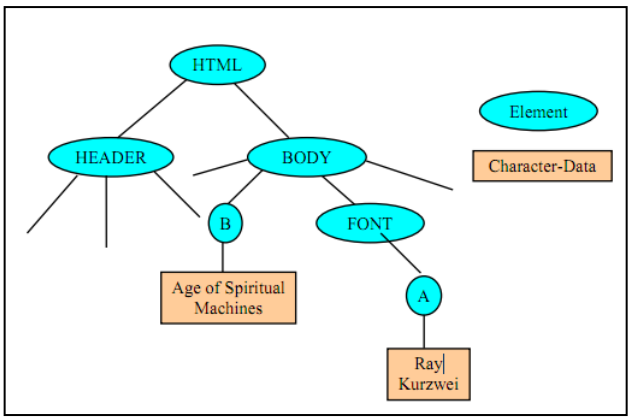
Theo W3C thì DOM (Document Object Model) [19] là một giao diện lập trình ứng dụng (API) cho các văn bản HTML hợp lệ và các văn bản XML có cấu trúc chặt chẽ.

Để trích xuất được thông tin cần thiết ở một node của cây DOM, chúng ta cần chỉ rõ đường đi từ gốc của cây đến node cần trích xuất thông tin. Đường đi này gọi là một Xpath hay mẫu trích xuất [18].

Các mẫu trích xuất có thể được làm rõ như đường dẫn từ gốc của cây DOM đến node chứa nội dung cần trích xuất.

Ví dụ:

Đây là cây DOM của một đoạn mã HTML chứa thông tin về cuốn sách, gồm tên cuốn sách (title) và tên tác giả (author). Bài toán đặt ra là sử dụng cây DOM này trích xuất các thông tin về tên sách và tác giả viết sách. Mẫu trích xuất được xây dựng sau:



Hình 2. *Mô hình DOM*

Mẫu trích xuất tên sách:

HTML→BODY→B→CharacterData.

Mẫu trích xuất tên tác giả:

HTML→BODY→FONT→A→ CharacterData.

c. Trích xuất thông tin dựa vào biểu thức chính quy

Với một biểu thức chính quy [20], một otomat hữu hạn trạng thái có thể được xây dựng và được sử dụng để so khớp sự xuất hiện của nó trong chuỗi tuần tự các trang web. Trong quá trình này, dữ liệu có thể được trích xuất.

Ví dụ: Với mã HTML như sau:

```
<head>
<meta http-equiv="Content-Language" content="en-us">
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<title>Tinh Tong cua cac so tu 1->n</title>
</head>
```

Để lấy được phần tiêu đề của đoạn mã này thì ta có thể xây dựng biểu thức chính quy như sau:

<head>.*?<title>(#text)</title>.

Giải pháp chúng tôi chọn thực hiện cũng dựa trên phương pháp bóc tách nội dung: sử dụng phương pháp tự động trong phân tích mã HTML để tạo thành cây Document Tree và biểu thức chính quy. Từ đó áp dụng các công cụ và kỹ thuật ngôn ngữ để quyết định phần nội dung chính. Tức là áp dụng phương pháp phân tích mã HTML kết hợp xử lý ngôn ngữ tự nhiên.

3.2. Phân loại thông tin

Phân loại văn bản tự động là lĩnh vực được chú ý trong những năm gần đây. Để phân loại người ta dựa trên nhiều cách tiếp cận khác nhau như dựa trên từ khóa, dựa trên ngữ nghĩa các từ có tần số xuất hiện cao, mô hình Maximum Entropy, tập thô, … Tiếng Anh là một trong những ngôn ngữ được nghiên cứu sớm và rộng rãi nhất với kết quả đạt được rất khả quan. Dựa trên các thống kê của Yang & Xiu (1999) và nghiên cứu của chúng tôi, một số phương pháp phân loại văn bản thông dụng có thể kể đến như: Support Vector Machine (SVM), K–Nearest Neighbor (kNN), Naïve Bayes (NB), Neural Network (NNet), Linear Least Square Fit (LLSF), Centroid- based vector,…

Các phương pháp phân loại văn bản đã được sử dụng thành công trên nhiều ngôn ngữ (Anh, Pháp, …). Tuy nhiên, trong tiếng Việt đơn vị nhỏ nhất là "tiếng" không phải là "từ" như trong tiếng Anh. Dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Tách từ trong tiếng Việt cũng là một thách thức khá thú vị.

3.2.1. Các phương pháp tách từ tiếng Việt hiện nay

Có rất nhiều những nghiên cứu trong tách từ tiếng Việt. Trong đó, nghiên cứu có độ chính xác cao nhất là của Lê Hồng Phương [8]: công cụ vnTokenizer cho kết quả chính xác tới 97,2%. Nghiên cứu này kết hợp các phương pháp máy hữu hạn trạng thái, phân tích dạng chính tắc, và ghép cực đại. Nhược điểm lớn nhất của phương pháp này là không xử lý được những từ mới. Như thế, phương pháp này không sử dụng kĩ thuật học máy.

Một số phương pháp tách từ tiếng Việt hiện nay có thể kể đến như: Maximum Matching, hình tách từ bằng WFST (Weighted Finite State Transducer) và mạng Neural, giải thuật học cải biến, quy hoạch động, tách từ dựa trên thống kê từ Internet và thuật toán di truyền,…

Với phương pháp Maximum Matching đã có nhiều nghiên cứu với kết quả thực nghiệm rất khả quan. Trong tiếng Trung, cách này đạt được độ chính xác 98,41% [Chih-HaoTsai, 2000]. Theo phương pháp LRMM để phân đoạn từ tiếng Việt trong một ngữ/câu, ta đi từ trái sang phải và chọn từ có nhiều âm tiết nhất mà có mặt trong từ điển, rồi cứ tiếp tục cho từ kế tiếp cho đến hết câu. Với cách này, ta dễ dàng tách được chính xác các ngữ/câu như: "hợp tác xã | mua bán"; "thành lập | nước | Việt Nam | dân chủ | cộng hoà",… Phương pháp RLMM thì ngược lại, trong một câu/ngữ, ta đi từ phải sang trái và chọn từ có nhiều âm tiết nhất mà có mặt trong từ điển, rồi cứ tiếp tục cho từ kế tiếp cho đến hết câu. Phương pháp MMSEG là sự kết hợp của cả hai phương pháp LRMM và RLMM, do đó MMSEG cho kết quả tốt hơn hai phương pháp trên.

Trong bài báo chúng tôi chọn phương pháp MMSEG [10] để tách từ tiếng việt, trong đó có sử dụng từ điển Tiếng Việt.

3.2.2. Áp dụng phương pháp máy học vector hỗ trợ SVM cho bài toán phân loại văn bản

Phương pháp SVM ra đời từ lý thuyết học thống kê và có nhiều tiềm năng phát triển trong thực tiễn.

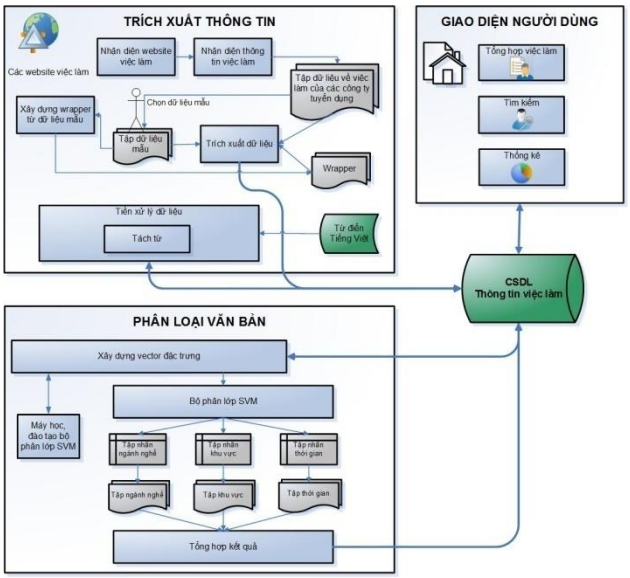
SVM có nhiều đặc tính nổi bật trong cả lý thuyết và thực tiễn so với các phương pháp khác trong lĩnh vực phân lớp văn bản.

Tuy rằng không gian vector đặc trưng ảnh hưởng rất lớn đến hiệu suất của phương pháp SVM. Nhưng trong bài báo của tôi chỉ phân loại thông tin theo 3 tiêu chí: ngành nghề, khu vực, thời gian. Ta nhận thấy tập từ vựng rút ra từ tập dữ liệu huấn luyện là hữu hạn, có kích thước nhỏ. Do đó, không gian vector đặc trưng sẽ không lớn và thời gian huấn luyện sẽ không nhiều, điều này khiến cho hiệu suất của phương pháp SVM [5] là tốt hơn.

Khó có thể trình bày một cách ngắn gọn và dễ hiểu trong một bài báo, nếu muốn hiểu rõ hơn thuật toán, bạn hãy đọc trực tiếp những thuật toán đó.

3.3. Đề xuất mô hình giải pháp

Mô hình giải pháp được mô tả như sau:



Hình 3. *Mô hình giải pháp*

Như mô hình trên, chúng tôi tóm tắt việc trích rút và tổng hợp thông tin việc làm bằng 6 bước như sau:

Bước 1: Nhận diện trang web việc làm

Từ rất nhiều website trên Internet ta tiến hành nhận diện đâu là website việc làm. Bước này ta có thể hình dung như là ta đang xây dựng một con robot có thể tự động dò tìm trong vô số website, sau đó nhận dạng đâu là website việc làm. Có thể thực hiện bằng cách đọc thông tin từ keyword trong metadata của trang web nếu nó có các từ khóa như: job, việc làm, tuyển dụng, công việc,... thì xác định đó là website việc làm.

Đầu vào: Các website trên Internet.

Đầu ra: Website việc làm.

Bước 2: Nhận diện thông tin việc làm

Từ website thông tin việc làm đã nhận dạng được ở Bước 1, ta tiến hành phân tích cấu trúc website nhận dạng đâu là thông tin việc làm, đâu là thông tin rác (tin tức, quảng cáo,…).

Đầu vào: Website việc làm.

Đầu ra: Thông tin việc làm.

Bước 3: Xây dựng wrapper từ dữ liệu mẫu

Từ những trang chứa thông tin việc làm, ta tiến hành chọn lựa những trang thông tin việc làm có cấu trúc chung nhất để làm dữ liệu mẫu cho việc xây dựng wrapper.

Đầu vào: Các trang thông tin việc làm.

Đầu ra: Tạo được wrapper từ dữ liệu mẫu đã chọn.

Bước 4: Trích xuất thông tin từ wrapper đã xây dựng

Từ wrapper đã xây dựng ở bước 3, ta tiến hành gán nhãn những kết quả đầu ra. Sau đó ta sử dụng wrapper đã được gán nhãn để thực hiện trích xuất thông tin thông việc làm.

Đầu vào: Wrapper đã xây dựng ở bước 3.

Đầu ra: Thông tin việc làm được trích xuất.

Bước 5: Tách từ tiếng Việt, phân loại thông tin

Ta tiến hành tách câu và tách từ cho mỗi văn bản dữ liệu việc làm. Không giống như trong tiếng Anh, việc tách từ trong văn bản tiếng Việt gặp nhiều khó khăn do sự phức tạp về cấu trúc cũng như đa dạng về ngữ nghĩa của từ tiếng Việt.

Đầu vào: Văn bản thông tin việc làm.

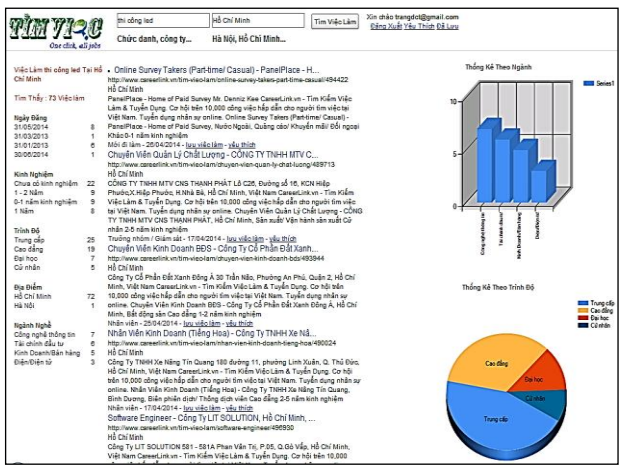
Đầu ra: Văn bản thông tin việc làm đã được phân loại theo các tiêu chí: khu vực, ngành nghề, thời gian.

Bước 6: Hiển thị, thống kê

Thông tin việc làm được hiển thị lại một cách tổng hợp và được phân loại theo các tiêu chí phục vụ cho việc tìm kiếm thuận tiện, dễ dàng. Ngoài ra, còn thống kê những số liệu (dưới dạng biểu đồ hình tròn hoặc hình trụ,…) phục vụ cho việc dự đoán xu hướng, định hướng và tư vấn việc làm.

4. Kết quả đạt được

Từ giải pháp như phân tích ở trên, chúng tôi tiến hành xây dựng hệ thống. Hệ thống này sử dụng ngôn ngữ ASP.NET kết hợp với hệ quản trị cơ sở dữ liệu SQL Server 2008. Kết quả trang thông tin tổng hợp về việc làm được tổng hợp từ các website việc làm như Hình 4.



Hình 4. *Kết quả trang thông tin tổng hợp việc làm*

Như hình trên, thông tin được tổng hợp từ các trang web khác nhau và được hiển thị lại một cách hệ thống, trực quan và có phân loại theo nhiều tiêu chí khác nhau nhằm giúp cho người dùng có thể dễ dàng tìm kiếm thông tin việc làm. Bên cạnh đó, hệ thống còn hiển thị số liệu thống kê việc làm theo thời gian, ngành nghề, địa bàn, yêu cầu về kinh nghiệm và trình độ. Các liệu thống kế này còn được biểu diễn dưới dạng biểu đồ rất trực quan nhằm giúp người dùng dễ dàng nhìn thấy một cách tổng quan về tình hình của thị trường lao động.

5. Kết luận

Nhu cầu về tìm kiếm và giới thiệu việc làm trực tuyến ngày càng được quan tâm đáng kể, đã kéo theo sự ra đời và tồn tại quá nhiều trang web về việc làm như hiện nay. Thực trạng này đã gây ra không ít khó khăn và lúng túng cho người sử dụng bởi lẽ thông tin về việc làm hiển thị rất rời rạc, rải

rác và đôi khi trùng lặp. Nhằm giải quyết vấn đề này, bài báo đã đề ra giải pháp biểu diễn thông tin việc làm một cách trực quan và hệ thống giúp người dùng dễ dàng tìm kiếm cũng như nhìn thấy bức tranh tổng thể hơn về việc làm. Với mục tiêu đó, chúng tôi đã xây dựng được hệ thống hỗ trợ tổng hợp và tư vấn thông tin việc làm với nhiều tính năng nổi bật như thu thập thông tin tự động, biểu diễn thông tin việc làm theo tiêu chí chọn lựa, thống kê số liệu việc làm theo lĩnh vực, trình độ, kinh nghiệm và thời gian.

Bài báo cũng đã trình bày việc nghiên cứu và ứng dụng các phương pháp trích rút tự động dựa vào phân tích mã HTML để tạo thành cây Document Tree và biểu thức chính quy để xây dựng hệ thống. Từ đó áp dụng các công cụ và kỹ thuật ngôn ngữ để quyết định phần nội dung chính. Phần phân loại thông tin sử dụng trong hệ thống là phương pháp máy học vector hỗ trợ SVM với việc áp dụng phương pháp tách từ tiếng Việt Maximum Matching để phân loại thông tin việc làm theo các tiêu chí: ngành nghề, khu vực, thời gian.

TÀI LIỆU THAM KHẢO

[1] Vũ Thanh Nguyên, Trang Nhật Quang, "Ứng dụng thuật toán phân lớp rút trích thông tin văn bản FSVM trên Internet", *Tạp chí Phát triển KH&CN*, Tập 12, Số 05 - 2009.

[2] Nguyễn Thị Trang, "Nghiên cứu các phương pháp trích rút văn bản từ trang web và ứng dụng", *Luận văn Thạc sĩ*, Học viện Công nghệ Bưu chính Viễn thông, 2013.

[3] Phạm Cẩm Vân, "Ứng dụng khai phá dữ liệu để tư vấn học tập tại trường Cao đẳng Kinh tế Kỹ thuật Quảng Nam", *Luận văn thạc Sĩ,* Đại học Đà Nẵng, 2012.

[4] Phạm Thanh Hùng, "Ứng dụng semantic web để phát triển hệ thống tư vấn học tập", *Luận văn thạc sĩ*, Đại học Đà Nẵng, 2011.

[5] Nguyễn Hải Minh, "Khai phá dữ liệu từ các mạng xã hội để khảo sát ý kiến của khách hàng đối với một sản phẩm thương mại điện tử", *Luận văn thạc sĩ*, Đại học Đà Nẵng, 2013.

[6] Vũ Tiến Thành, "Bài toán trích xuất thông tin cho dữ liệu bán cấu trúc và áp dụng xây dựng hệ thống tìm kiếm giá cả sản phẩm*", Khóa luận tốt nghiệp,* Đại học Công nghệ, 2009.

[7] Trần Thị Thu Thảo, Vũ Thị Chinh, "Xây dựng hệ thống phân loại tài liệu tiếng Việt", *Nghiên cứu Khoa học*, 2012

[8] Le, H.P, Nguyen, T.M.H, Azim Roussanaly, Ho, T.V, A hybrid approach to Word Segmentaion of Vietnamese texts (2008), Language and atutomata theory and applications 2nd international coference, LATA 2008.

[9] Luu, T. A, Yamamoto, K., A pointwise approach for Vietnamese Diacritics Restoration, IALP 2012

[10] Chih-Hao Tsai, "MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm"

[11] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques, 2011

[12] Yongjian Fu, Data mining: Tasks, Techniques, and Applications, University of Rolla.

[13] T. Mitchell, Machine Learning and Data Mining, *Communications of the ACM*, Vol. 42 (1999), No. 11, pp. 30--36.

[14] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy: Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, (1996).

[15] V. Crescenzi, G. Mecca, and P. Merialdo. Roadrunner: Towards Automatic Data Extraction from Large Web Sites. In Proc. of Very Large Data Bases (VLDB'01), pp.109–118, 2001.

[16] http://nhuthuan.blogspot.com/2006/11/s-lc-v-k-thut-trong-vietspider-3.htm

[17] http://www.dcs.bbk.ac.uk/~ptw/teaching/ssd/slide2.html

[18] http://www.w3.org/TR/xpath

[19] http://www.w3.org/DOM

[20] http://www.cs.uic.edu/~liub/WebMiningBook.html

[21] http://www.gso.gov.vn

(BBT nhận bài: 18/09/2014, phản biện xong: 03/10/2014)